AFRL-RI-RS-TR-2009-316
Final Technical Report
January 2010

LEARNING COMPOSITIONAL SIMULATION MODELS

University of Massachusetts

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

NOTICE AND SIGNATURE PAGE

AFRL-RI-RS-TR-2009-316 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/s/	/s/
NANCY A. ROBERTS	JOSEPH CAMERA, Chief
Work Unit Manager	Information & Intelligence Exploitation Division
	Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (*DD-MM-YYYY*)	2. REPORT TYPE	3. DATES COVERED (*From - To*)
JANUARY 2010	Final	June 2007 – July 2009

4. TITLE AND SUBTITLE
LEARNING COMPOSITIONAL SIMULATION MODELS

5a. CONTRACT NUMBER N/A
5b. GRANT NUMBER FA8750-07-2-0158
5c. PROGRAM ELEMENT NUMBER N/A

6. AUTHOR(S)
David Jensen

5d. PROJECT NUMBER PAIN
5e. TASK NUMBER 00
5f. WORK UNIT NUMBER 02

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Massachusetts
70 Butterfield Terrace
Amherst MA 01003-9242

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFRL/RIED
525 Brooks Rd.
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
N/A

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TR-2009-316

12. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PA# 88ABW-2009-5081

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Effective and proactive decisions about intelligence gathering depend on accurate models of an adversary. Specifically, such models need to accurately reflect the cause-and-effect dependencies within the systemic behavior of the adversary. Such models can be created based entirely on the knowledge of experts, or they can be created or augmented based on the analysis of data. However, creating causal models from data will require advances in the fundamental science and technology of discovering causal knowledge. Our project focused on creating such advances. Specifically, we focused on automating the application of quasi-experimental designs, a set of manual analysis techniques developed by social scientists, economists, and medical researchers over the past four decades. Quasi-experimental designs (QEDs) are templates for causal discovery from observational (non-experimental) data. QEDs identify naturally occurring experiments that support inferences about causal dependencies within large bodies of observational data. Our work has shown that many potential designs exist for realistic tasks that those designs can increase the accuracy with which causal inferences can be made from small amounts of data, and that such designs can be automatically identified. This lays the groundwork for powerful tools with which analysts can examine observational data of complex organizations and system to improve their causal understanding of those systems.

15. SUBJECT TERMS
Quasi-experimental designs, statistical relational learning, causal discovery, machine learning, causal models

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
U	U	U	UU	21	Nancy A. Roberts
					19b. TELEPHONE NUMBER (*Include area code*) N/A

Table of Contents

Summary 1
Introduction ……………………………………………………………………………….2
Why Causal Models are Useful 2
Quasi-Experimental Design 4
Methods, Assumptions, and Procedures 7
Results and Discussion 8
Quasi-experimental design 8
Data 8
Findings 11
Relational learning 12
Conclusions 14
References 15
List of Acronyms 16

List of Figures

Figure 1: Three causal models that produce statistical association. 3
Figure 2: Example results from twin studies. 5
Figure 3: Graphical models representing monozygotic and dizygotic twins................................. 6
Figure 4: Entity-Relationship diagram for the IMDb+Netflix database.. 9
Figure 5: A schema for the Wikipedia data set... 10

Summary

We made substantial progress in two areas: (1) extending the expressiveness and ability to learn models of relational and temporal data; and (2) using those expressive representations to learn statistical models that express causal dependencies. The approach to learning causal models — exploiting what statisticians call quasi-experimental designs — is particularly promising, though this was only a preliminary study. Overall, the results indicate that substantial more work is warranted in automatic application of quasi-experimental designs. The work reported here resolves several questions that would otherwise have indicated that the approach was unlikely to have worked well.

Introduction

Recent advances in machine learning (ML) in complex data sets have revealed a surprising new opportunity to learn causal models of complex systems. The opportunity is a deep and unexploited technical interaction between two previously unconnected areas: (1) work in statistical relational learning; and (2) work on quasi-experimental design in the social sciences. Specifically, the type of new data representations conceived and exploited recently by researchers in statistical relational learning (SRL) may provide all the information needed to automatically apply powerful statistical techniques from the social sciences known as quasi-experimental design (QED). QEDs allow a researcher to exploit unique characteristics of sub-populations of data to make strong inferences about cause-and-effect dependencies that would otherwise be undetectable. Such causal dependencies infer whether manipulating one variable will affect the value of another variable, and they make such inferences based on non-experimental data.

To date, QEDs have been painstakingly applied by social scientists in an entirely manual way. However, data representations from SRL that record relations (organizational, temporal, spatial, and others) could facilitate automatic application of QEDs. Constructing methods that automatically identify sub-populations of data that meet the requirements of specific QEDs would enable powerful and automatic causal inferences from non-experimental data. This fusion of work in SRL and QED would lead to: (1) large increases in the percentage of causal dependencies that can be accurately inferred from non-experimental data; (2) large reductions in the amount of data needed to discover causal dependencies that can already be inferred; and (3) large reductions in the computational complexity of causal learning algorithms.

If exploited, this capability could provide a dramatic leap in the ability of intelligence analysts and others to automatically construct causal models of large and complicated systems (e.g., social systems, organizations, and computer systems). Such models would be a significant improvement over existing models learned by statistical and machine learning techniques, the vast majority of which are non-causal (and thus do not allow analysts to correctly infer the effects of potential actions) or only weakly causal (because many of the potential causal dependencies cannot be correctly inferred).

The goal of this research was to investigate the potential of this interaction and make fundamental advances in the techniques to exploit it. It should be noted that this is fundamental research that investigates new foundations for knowledge discovery algorithms rather than mere improvements to existing algorithms. While we produced significant advances in the shortened period of this contract, continued support in the future could lead to dramatic improvements the basic technologies and to an applied system.

Why Causal Models are Useful

Nearly all algorithms for machine learning analyze data to identify statistical associations among variables. That is, they identify variables of some entity (e.g., a patient's occupation, recent physical contacts, and symptoms) that are statistically associated with other variables (e.g., a disease). Such associations are useful for making predictions about the values of unobserved variables based on the values of variables that can be observed. For example, a doctor could predict whether a patient has a particular disease (an unobserved variable) based on a set of observed symptoms.

Such associational models can be useful in many situations. For example, associational models constructed by machine learning algorithms now sit at the heart of most state-of-the-art systems for machine translation, speech understanding, computer vision, information extraction, and information retrieval. In all of these cases, associations among variables alone are sufficient to meet the goals of the deployed system.

However, machine learning algorithms are often deployed in the hope that they will support decisions about which actions, or interventions, to make in a given situation. In the case of medical diagnosis, most medical professionals do not simply want to diagnose disease, but to prevent, treat, or mitigate the effects of the disease as well. They want to know what effect a particular intervention (e.g., implementation of a public health measure or widespread administration of a drug) will have on the health of a population. In such situations, practitioners want models that help them to design effective interventions, and this requires the modeling of causality, not merely statistical association.

Remarkably, most existing probabilistic models are practically useless for designing effective interventions because they only identify statistical associations, not causal dependencies. As is emphasized in nearly all introductory statistics courses, correlation is not causation — statistical association between two variables does not necessarily imply that one causes the other. For example, suppose we gathered a sample of patients and measured a variety of variables about each patient, including their history of smoking and their incidence of lung cancer.

If we analyzed the data, we would very likely find a statistical association between several of these variables.

However, the association between any two variables A and B could result from any of three causal situations shown in Figure 1. If A and B are associated, then A could cause B (smoking causes lung cancer), B could cause A (a predisposition to nicotine addiction causes smoking), or a third variable C could cause both A and B (genetics could cause both a predisposition to nicotine addiction and a predisposition to lung cancer).

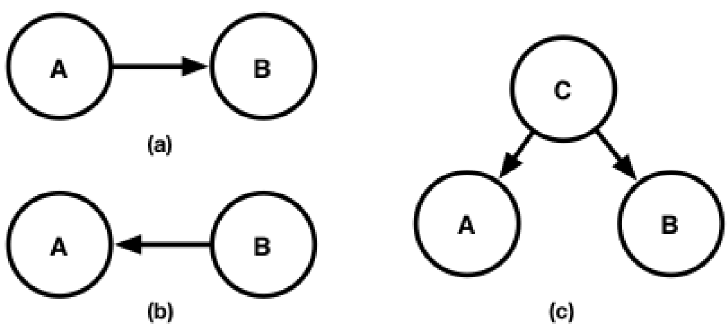

Figure 1: Three causal models that produce statistical association.

If a purely associational model is constructed from data and then used to support the design of interventions, either of the latter two cases could cause the resulting interventions to be ineffective (or even counterproductive).

In contrast, if accurate causal models could be constructed, they would be useful to a wide range of users within the Intelligence Community.

Quasi-Experimental Design

Fortunately, a class of methods does exist that can infer causal knowledge from observational data. These methods are routinely used to support causal inferences in medicine, economics, and social science. They can be grouped under the rubric “quasi-experimental design” (QED), and they attempt to exploit inherent characteristics of observational data sets that partially emulate the control and randomization possible in an experimental setting (Campbell, Stanley & Gage 1963; Cook & Campbell 1979; Shadish, Cook & Campbell 2002).

Although QEDs clearly do not always have the internal validity of traditional experimental designs, but they can be applied to the much wider array of data sets that modern data collection practices have made available, and the size and scope of those data sets can partially or completely compensate for the deficiencies that arise from lack of experimental control. Indeed, there are a wide variety of situations where causality can be explored in no other way.

In the absence of explicit control and randomization, some QEDs employ case matching to identify pairs of data instances that are as similar as possible in all respects except for the variable under investigation (the non-equivalent group design). Other QEDs examine how the value of a given variable on the same data instances changes over time, typically before and after some specific event (the regression-discontinuity design). Other types of quasi-experimental designs that have been devised include the proxy pretest design, double pretest design, nonequivalent dependent variables design, pattern matching design, and the regression point displacement design.

A particularly salient example of quasi-experimental design is a classical twin study, a design that has been employed for decades to study the causal factors for particular diseases and conditions. Twin studies compare the incidence of disease in sets of monozygotic (identical) and dizygotic (fraternal) twins. Monozygotic twins share identical genetics, a common fetal environment, and (typically) a common post-natal environment. The same is true for dizygotic twins, except that they are only genetically similar rather than genetically identical.

This remarkable degree of shared background, as well as the specific difference in the shared background between the two types of twins, provides a nearly ideal setting to study the effect of genetics on disease. For example, to identify the degree to which a given condition is due to genetic factors, investigators can determine the correlation in the condition among pairs of each type of twin, and then compare the correlation between the two types. A large difference indicates that a large portion of the condition is due to genetics, whereas no difference indicates that the condition is due to other factors. Figure 2 summarizes a few results from twin studies of various conditions.

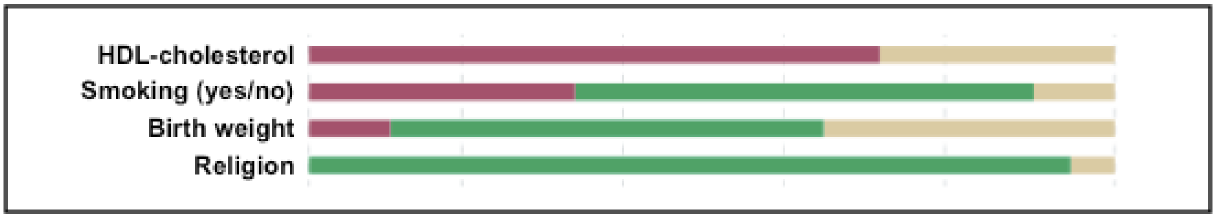

Figure 2: Example results from twin studies, drawn from a recent review (Boomsma, Busjahn & Peltonen 2002). Purple (darkest) bars indicate effects due to genetics, green (dark) bars to shared environment, and beige (light) bars to unique environment.

Two factors are remarkable about the efficacy of twin studies. First, they allow the quantitative impact of genetics to be determined even though investigators may have no idea what specific genes are involved. That is, they can determine the degree to which some variable on a particularly entity (genotype) affects the observed condition without knowing what that variable is or how to measure it. Second, they can perform this analysis by studying only a tiny fraction of an entire population. Indeed, without access to a very large population, it would be virtually impossible to gain access to a sufficient number of pairs of monozygotic and dizygotic twins.

It is also important to note that the validity of twin studies relies on at least three pieces of information known to investigators but not (typically) represented explicitly in data used for QED studies. First, twins occur relatively randomly in the population. If identical twins were much more likely to be born to parents with particular genetic traits or who lived in particular environments, then those factors would confound efforts to use twins to study the effects of genetics on physical conditions. Second, genetic makeup is established temporally prior to the onset of diseases and other conditions. Thus, we know the direction of causality without having to determine it from data.

Finally, and perhaps most importantly, we know that genotype (genetic sequence) and phenotype (physical condition) can be treated as related but separate entities. This means that individuals can have identical genotypes but not identical phenotypes. Thus, the relational representation shown in Figure 3 can be used to represent the data, where monozygotic twins share a common genotype and dizygotic twins do not. This relational representation underlies the inference of investigators that, if the two types of twins do not differ significantly in the correlation of their conditions, then all possible variables on genotype can be removed as potential causal factors.

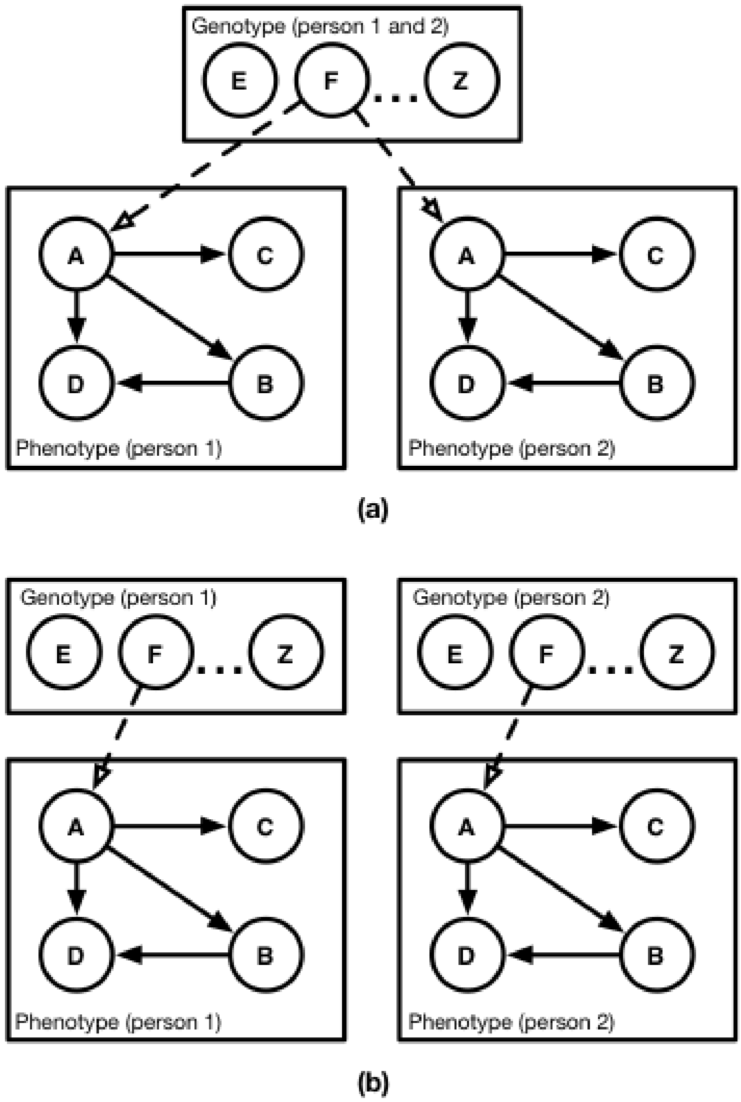

Figure 3: Graphical models representing monozygotic (a) and dizygotic (b) twins. Circles represent variables, boxes represent entities, and solid and dashed arrows represent known and potential dependencies, respectively.

Methods, Assumptions, and Procedures

The early goals of our study (which were the only ones accomplished due to early termination of the cooperative agreement) were to assess:

(1) *Applicability* — Determine the degree to which QEDs could be used to address causal questions of interest;

(2) *Utility* — Assess the qualitative and quantitative impact of using QEDs for causal discovery; and

(3) *Potential for automated identification* — Determine whether it was possible to identify QEDs automatically.

Our methods included the following:

(1) *Literature review* — We reviewed key texts on quasi-experimental design (Campbell & Stanley 1963; Cook & Campbell 1979; Shadish, Cook & Campbell 2002), major journal articles, and websites devoted to methods for quasi-experimental design. This work required substantial translation between concepts common to machine learning and those common in social science and experimental design. For example, much of the research in QEDs solidified before the revolution in ML that led to the widespread adoption of graphical models, and this necessitated a translation of QED concepts into the terms of graphical models.

(2) *Construction of test problems* — We gathered data and created the necessary background knowledge for four test domains that supported both manual and automated analysis. Three of the domains were drawn from real domains for which data were available (Wikipedia, the National Football League, and the US motion picture industry) and one was drawn from a realistic domain for which we lack data (military flight training). Our criteria for selecting domains included: (a) the existence of a rich relational structure that supports the identification of QEDs; (b) the similarity to military scenarios involving both collaboration and adversarial behavior; (c) the likelihood of intrinsic interest among other machine learning researchers; (d) the existence of causal intuitions on the part of the PI and graduate students; and (e) the availability of real data.

(3) *Manual identification of QEDs* — We manually identified QEDs in each of the test domains in order to gain experience with the necessary knowledge representation and reasoning capabilities.

(4) *System construction* — We encoded some key concepts of quasi-experimental design into first-order logic and constructed a prototype for identifying simple QEDs by automated reasoning using those logical rules.

(5) *Evaluation* — We performed an initial evaluation on that prototype.

Results and Discussion

The results of our work fall into two broad categories: (1) exploratory research on the potential to use quasi-experimental designs to learn causal models; and (2) more traditional research on developing representations and algorithms for learning highly expressive models of complex types of data. Prior work in the second area has created the opportunity explored in the first area, and continued work in the second area (expressing and learning highly expressive statistical models) is necessary for full exploitation of the first area.

Quasi-experimental design

The results of our work on quasi-experimental designs are preliminary, largely due to the early termination of the cooperative agreement. However, based on our preliminary work, the prospects for using quasi-experimental designs to significantly enhance causal discovery is excellent.

Data

As already mentioned, we assembled and analyzed several data sets to provide experience and case studies concerning the applicability and variety of quasi-experimental designs. In particular, we assembled data from the National Football League, the Internet Movie Database, and Wikipedia.

National Football League — The National Football League (NFL) is the governing body for American football franchises. The league consists of 32 teams that are divided into two conferences of equal size — the American Football Conference (AFC) and National Football Conference (NFC). Each conference has four equal-sized divisions of four teams each. Each team participates in 16 regular season games per year. Six of those games are with the other three divisional rivals (once at home and once away). We gathered data over a five-season span (2002-2007), drawing in particular from one online source,[1] though many sources provide similar data.

American football, and sports in general, are a rich tradition of causal questions. For example, does playing at particular stadiums cause relative score differentials for the home team? Does fatigue induced by multiple away games produce lower scores? We used QEDs enabled by the structure of the NFL data to examine several of these questions. Such designs are enabled by the structure of the NFL data. For example, the regular structure of divisional games (playing the same team once at home and once away) provides a nearly perfect example of a blocking design in which each of a set of entities (each of a set of team pairings, in this case) is subjected to two different treatments. Blocking designs help control for both measured and unmeasured variables on those entities, because blocking designs examine the differential effect of treatments when all those variables are held constant.

Internet Movie Database — The Internet Movie Database[2] contains information on movies released worldwide, including release dates, directors, producers and actors, as well as the nominees and recipients of Academy Awards. We selected a subset of these awards covering films released in the years 1997 to 2007. We included information on the nominees and winners

[1] www.pro-football-reference.com
[2] www.imdb.com

of Best Picture, Best Director, Best Actor, and Best Actress. We augmented the IMDb data with the Netflix Prize data set,[3] which contains the title and year of release for 17,770 movies released on DVD and ratings of those movies from more than 400,000 customers. The date range for ratings is from November 11, 1999 to December 31, 2005. The schema shown in figure Figure 4: Entity-Relationship diagram with temporal frequencies and extents for the IMDb+Netflix database. Each movie has a series of actor and director stints as well as a review by a user of the Netflix Prize database. Awards are presented to actors, directors, and movies. represents the combination of the two data sets.

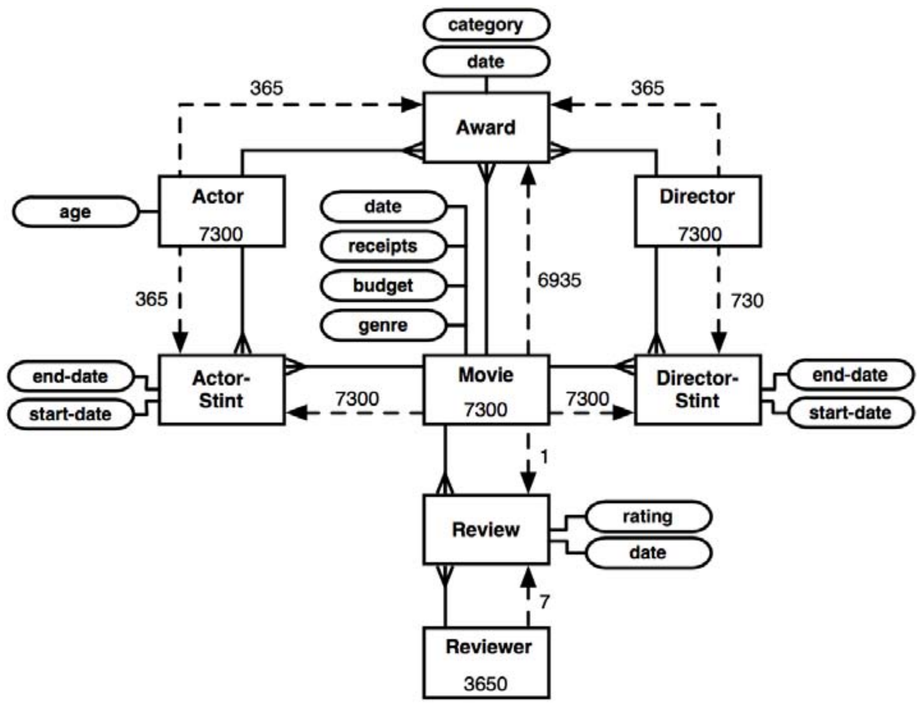

Figure 4: Entity-Relationship diagram with temporal frequencies and extents for the IMDb+Netflix database. Each movie has a series of actor and director stints as well as a review by a user of the Netflix Prize database. Awards are presented to actors, directors, and movies.

The data could be used to examine a large number of interesting causal questions, and the structure of the data provides a large number of potential designs. For example, one instance of a QED identified by automatically by our prototype AIQ system involves the variables of award existence and an aggregate of user ratings on a base item of movies. This design implies, rather intuitively, that granting an Academy Award to a movie may cause changes in user ratings of that movie. This design was made possible because whether an award entity exists was designated as pseudo-random among all nominated movies (i.e., all nominated movies are equally likely to win an award). This is clearly an assumption, but a plausible one.

We tested this design by computing the average rating a movie receives in the two months prior to and the two months after Academy Awards are granted. For each movie, we computed the difference in the average ratings. Then we compared the mean difference for movies that won an

[3] www.netflixprize.com

award with the mean difference for those who were nominated but did not win. The difference was found to be weakly significant, implying a causal connection between whether a movie wins an Academy Award and how Netflix viewers rate that movie.

Importantly, this design avoids the obvious problems with a more simplistic analysis that would merely compare the Netflix ratings of all movies that won awards with the ratings of all movies that did not. The results of such an analysis (winners are more highly rated) could be due to the fact that a third variable (movie quality) is a common cause of both winning awards and receiving high ratings. Additional details can be found in our KDD 2008 paper (Jensen et al. 2008).

Wikipedia — Wikipedia is a peer-produced general knowledge encyclopedia.[4] Wikipedia articles, or pages, are produced collectively by thousands of volunteer users. Because the articles are created, modified, stored, and read entirely in an online environment, and because the users and editors of Wikipedia are geographically dispersed, nearly all interactions with the system and between users are entirely captured by Wikipedia's logs.

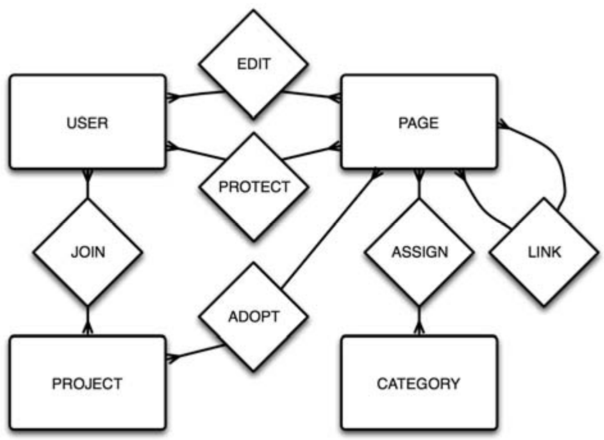

Figure 5: A schema for the Wikipedia data set, showing how entities of users, pages, projects, and page categories are related.

Figure Figure 5: A schema for the Wikipedia data set, showing how entities of users, pages, projects, and page categories are related. provides a simplified relational data schema that describes the major entities and relations that make up Wikipedia. Pages are created and modified by users, and users often organize themselves into groups called projects, each of which covers a general topic. Within a project, editors assess individual pages for "importance" (how central the page is to the project theme) and "quality" (a project-independent objective evaluation of key criteria).

The data provide the ability to examine many causal claims about Wikipedia. For example, one of the most persistent claims about Wikipedia is that its reputability stems from the large number of users that collaborate to write each article. We call this the "many-eyes hypothesis"—the more users that revise an article, the higher the quality of that article. If we knew that this claim

[4] www.wikipedia.org

were actually causal, then we could theoretically increase the quality of an article by asking more users to participate in revisions.

However, to actually determine that there exists a causal dependence between the number of users editing an article and its quality, we must eliminate other plausible alternative models that could explain the observed correlation. Fortunately, the data available on Wikipedia make it possible to evaluate this claim. In fact, the data allow the use of a number of different designs, each eliminating different potential threats to a valid causal conclusion. Other interesting causal questions include whether page adoption by a project increases page quality, what effects vandalism has on the frequency with which a page is monitored and edited, and whether joining a project increases a user's participation in editing. Additional details can be found in a recent technical report (Maier, Rattigan, and Jensen 2009).

Findings

Our analysis of specific case studies in the data above established the following findings:

- *QEDs are widely applicable* — For the domains examined in the study, many causal questions could be examined with one or more quasi-experimental designs. Such QEDs were not always immediately apparent, but this parallels findings of several studies of the use of quasi-experimental designs in the social sciences, which indicate that relatively simple designs are often used when alternative designs with higher statistical power are available.
- *QEDs are useful in key domains*— In several cases, using quasi-experimental designs allowed the elimination of one or more potential causal models, thus restricting the space of models that can account for the observed correlations. Quantifying the extent to which designs increase the accuracy and statistical power of causal inference was not possible at this early stage, but the utility appears relatively large.
- *QEDs can be automatically identified* — At least two simple designs (designs that use quasi-control and designs that employ blocking) can be identified automatically using a relatively simple automated reasoning engine. This automated identification requires reasoning systems that represent and reason about ontological information and known constraints on the causal model. If such information is represented, then a reasoning system using first-order logic can identify at least some quasi-experimental designs (Jensen et al. 2008).

We developed the following capabilities:

- *An automated reasoning system for identifying one type of QED* — We developed the AIQ system (Automated Identification of Quasi-experiments) that can automatically identify and reason about designs that rely on quasi-random variables (Jensen et al. 2008). The system takes as input a data schema and an existing set of causal knowledge and produces as output one or more QEDs, where each QED includes a treatment variable (the potential cause), an outcome variable (the potential effect), and a specification of a unity (the portion of the data schema used to create the rows of a data table to be used for hypothesis testing). Quasi-random variables are essentially the simplest of all QEDs, but the necessary infrastructure to identify designs was non-trivial. AIQ is written in Prolog and is available as open-source software.
- *Evaluation data sets and case studies* — We developed several data sets and case studies to both evaluate and demonstrate AIQ. As mentioned above, three of the domains were drawn

from real domains for which data were available (Wikipedia, the National Football League, and the US motion picture industry) and one was drawn from a realistic domain for which we lack data (military flight training).

Relational learning

Our work in relational learning follows on from a well-established line of research over the past decade. Our work under this contract resulted in the following capabilities:

- *An error analysis framework for relational models* — We developed a bias-variance framework for relational models that decomposes loss into errors due to both the relational learning and the collective inference processes (Neville & Jensen 2008). We evaluated the performance of three relational models on synthetic and real-world datasets with the framework and showed that: (1) inference can be a significant source of error; and (2) the models exhibit different types of errors as data characteristics are varied.
- *A framework for learning temporal-relational models* — We developed a novel framework for learning predictive models of dynamic relational data (where the relationships among entities change over time) (Sharan & Neville 2008). We use a two-phase process that first summarizes the temporal changes in link structure by constructing a static, weighted relational graph using kernel smoothing and then we learn modified statistical models from the summarized data. This approach facilitates the development of efficient learning and inference techniques by considering a restricted set of temporal-relational dependencies and using parameter-tying methods to generalize across relationships and entities.
- *A method for resampling from relational data* — We developed a relational resampling technique to accurately estimate the variance of sampling distributions of statistics for heterogeneous, dependent data (Eldardiry & Neville 2008). The approach aims to preserve local relational dependencies (e.g., relational autocorrelation) and link structure, while introducing sufficient variance at a global level to correctly model the process of drawing samples from the underlying population. The key idea behind the approach is to sample subgraphs with replacement from the original data, thereby preserving the local link and attribute structure within the subgraphs. This is augmented with a procedure that links up the selected subgraphs in an attempt to match the global properties of the data without reproducing them exactly.
- *A new learning method for within-network classification* — We developed a categorization framework for "within-network" relational learning and inference, where models are learned on a partially labeled relational dataset and then are applied to predict the classes of unlabeled instance in the same graph (Xiang & Neville 2008). Current methods can be categorized as: disjoint learning with disjoint inference, disjoint learning with collective inference, and collective learning with collective inference. Here "disjoint" refers to techniques that ignore the unlabeled data and "collective" refers to techniques that jointly consider the labeled and unlabeled data. Models from each of these categories have been employed previously in different settings, but to our knowledge there has been no systematic investigation comparing models from the three categories simultaneously. To undertake this investigation, we developed a novel pseudolikelihood EM method that facilitates collective learning and collective inference on partially labeled relational networks. We then compare

this method to competing methods drawn from the same family of models to investigate the performance tradeoffs between disjoint and collective modeling approaches.

- *A shrinkage approach to model non-stationary relational dependencies* — Current statistical relational learning techniques model *global* autocorrelation dependencies under the assumption that the level of autocorrelation is stationary throughout the graph (Angin & Neville 2008). To date, there has been no work examining the appropriateness of this stationarity assumption. We examined two real-world datasets and found that there is significant variance in the autocorrelation dependencies throughout relational data graphs. To account for this, we developed three shrinkage techniques for modeling non-stationary autocorrelation, which combine local and global estimates of the relational dependencies in the data. In regions of the graph where there is sufficient information locally for accurate parameter estimation, the model relies on the local estimates more heavily; otherwise it backs off to the global estimates. This results in a modeling approach that is more robust to variance in relational dependencies throughout a relational data graph.
- *A link-strength prediction method that uses transactional information among entities* — In electronically collected social networks data sets, the observed relationships often contain noisy information (i.e., weak relationships) (Kahanda & Neville 2009). Since the accuracy of relational modeling techniques is often contingent on the presence of links in the data that confer homophily, methods that can prune away these spurious relationships and highlight stronger relationships will likely result in improved model performance. Online social network domains contain ancillary transactional data among entities (e.g., communication, file transfers) that can be used to infer the true underlying social network. We exploit this transactional information and developed models to predict "link strength" based on topological, and transactional features. We evaluated our approach on real-world data and showed that we can accurately predict strong relationships. Moreover, we show that transactional-network features are the most influential features for this task.

Conclusions

In addition to the above results, we conclude:

(1) *More designs exist than can be easily handled manually* — Many designs exist for any one causal inference task. For a large number of tasks we considered manually, the number of possible designs quickly exceeded our capacity to easily consider them without automated assistance. This has positive implications for the prospect of automated systems to help human analysts simultaneously reason about the space of causal models and designs.

(2) *Blocking is a highly useful design element* — While we began the study expecting to implement entire QEDs, we quickly came to realize that many designs that are implemented by human experts combine multiple design *elements* (e.g., temporal blocking, modeling, and multiple pre-tests and post-tests). Among these options, *blocking* is among the most widely useful elements of QEDs. Blocking groups data elements into "blocks" within which many variables are controlled. For example, blocks could be pairs of twins (who share a common genome), groups of employees in a given company (who share a common work environment), or even a single person at two different times (where variables about that person are assumed to remain constant over the entire time period).

(3) *QEDs can increase power* — QEDs can substantially increase the statistical power of a system, allowing it to make causal inference with far fewer data points than would otherwise be possible. Among the ways in which they do this is to reduce the need to model (or even measure) some set of variables. Fewer variables means lower sample complexity (increased power).

(4) *QEDs can interact to allow chains of causal inference* — Interactions occur between conclusions of individual designs, and these interactions could produce a beneficial "chain reaction" when they are applied sequentially. This has strong implication for the design of systems that exploit QEDs, because it implies that some designs are more useful to apply first because the conclusions they support can enable other designs.

(5) *Relational models require specialized evaluation methods* — Relational models require fundamentally different types of analysis than propositional models, as we detail in a recent paper (Neville & Jensen 2008).

(6) *Representational expressiveness is still a limiting factor in model accuracy* — Our work on temporal-relational models and modeling non-stationary autocorrelation shows that significant improvements in accuracy can be obtained by increasing the expressiveness of the space of models considered by the learning algorithm. That is, for many data sets of realistic size, the expressiveness of the model space is still a limiting factor on the accuracy of the models that can be learned. This has significant implications for the prospects for learning causal models. Learning models of non-trivial domains, and applying the full range of quasi-experimental designs, will require models that go beyond simple relational representations to include non-stationary distributions and temporal dependencies.

References

Angin, P. and J. Neville (2008). A shrinkage approach for modeling non-stationary relational autocorrelation. *Proceedings of the 2008 IEEE International Conference on Data Mining (ICDM)*.

Boomsma, D., Busjahn, A., and Peltonen, L. (2002). Classical twin studies and beyond. *Nature Reviews Genetics* 3:872-882

Campbell, D., J. Stanley, and N. Gage (1963). *Experimental and Quasi-experimental Designs for Research.* Rand McNally.

Cook, T., and D. Campbell (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Wadsworth.

Eldardiry, H. and J. Neville (2008). A resampling technique for relational data graphs. *Proceedings of the 2nd Workshop on Social Network Mining and Analysis at the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Jensen, D., A. Fast, B. Taylor, and M. Maier (2008). Automatic identification of quasi-experimental designs for discovering causal knowledge. *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Kahanda, I. and J. Neville (2009). Using transactional information to predict link strength in online social networks. *Proceedings of the 2009 AAAI International Conference on weblogs and Social Media (ICWSM).*

Maier, M., M. Rattigan, and D. Jensen (2009). Discovering causal knowledge by design. Department of Computer Science. University of Massachusetts Amherst. Technical Report UM-CS-2009-047.

Neville, J. and D. Jensen (2008). A bias/variance decomposition for models using collective inference. *Machine Learning Journal* 73:87-106.

Shadish, W., T. Cook and D. Campbell (2002). *Experimental and Quasi-Experimental Designs*. Houghton Mifflin.

Sharan, U. and J. Neville (2008). A framework for exploiting temporal variations in relational domains. *Proceedings of the 2008 IEEE International Conference on Data Mining (ICDM).*

Xiang, R. and J. Neville (2008). Pseudolikelihood EM for within-network relational learning. *Proceedings of the 2008 IEEE International Conference on Data Mining (ICDM).*

List of Acronyms

ML – Machine Learning

PAINT – Proactive Intelligence

QEDs – Quasi-experimental Designs

SRL – Statistical Relational Learning